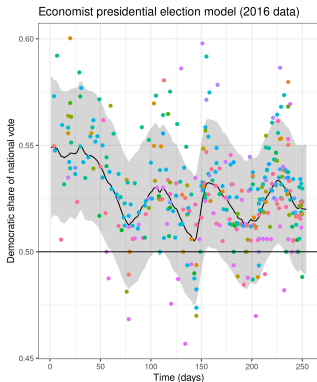


Approximate data deletion and replication with the Bayesian influence function

Ryan Giordano (rgiordano@berkeley.edu, UC Berkeley), Tamara Broderick (MIT)

Theory and Foundations of Statistics in the Era of Big Data — Honoring Basu and Bahadur (April 2024)

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

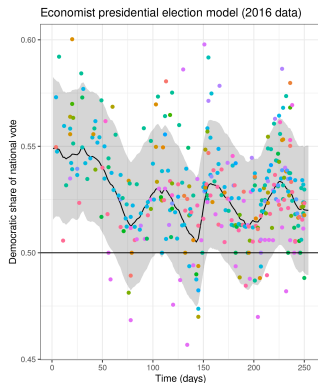
Model:

- $X = x_1, \dots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\mathbb{E}_{p(\theta|X)} [f(\theta)]$.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \dots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

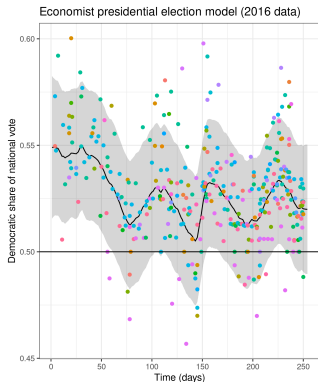
We want to know $\mathbb{E}_{p(\theta|X)} [f(\theta)]$.

The people who responded to the polls were randomly selected.

If we had selected a different random sample, how much would our estimate have changed?

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \dots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\mathbb{E}_{p(\theta|X)} [f(\theta)]$.

The people who responded to the polls were randomly selected.

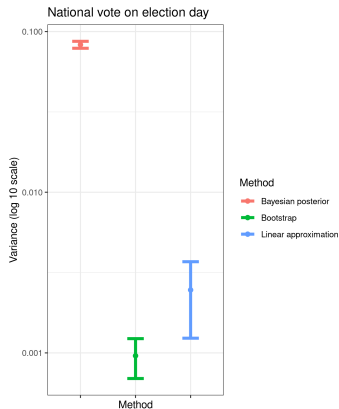
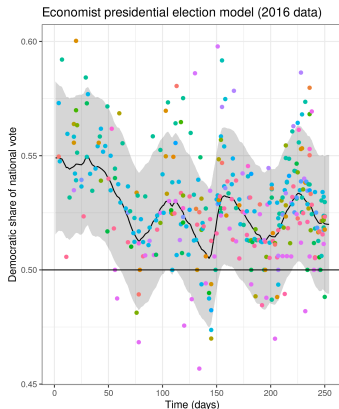
If we had selected a different random sample, how much would our estimate have changed?

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Problem: Each MCMC run takes about 10 hours (Stan, six cores).

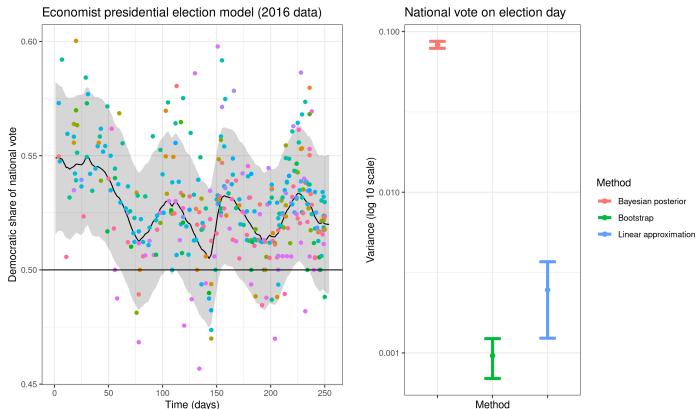
Proposal: Use full-data posterior draws to form a linear approximation to *data reweightings*.

Proposal: Use full-data posterior draws to form a linear approximation to *data reweightings*.



Results

Proposal: Use full-data posterior draws to form a linear approximation to *data reweightings*.



Compute time for 100 bootstraps: 51 days

Compute time for the linear approximation: Seconds
(But note the approximation has some error)

- Data reweighting
 - Write the change in the posterior expectation as **linear component** + **error**
 - The **linear component** can be computed from a single run of MCMC

- Data reweighting
 - Write the change in the posterior expectation as **linear component** + **error**
 - The **linear component** can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
 - As $N \rightarrow \infty$, the linear component provides an arbitrarily good approximation

- Data reweighting
 - Write the change in the posterior expectation as **linear component** + **error**
 - The **linear component** can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
 - As $N \rightarrow \infty$, the linear component provides an arbitrarily good approximation
- High-dimensional problems
 - The linear component is the same order as the error
 - Even for parameters which concentrate, even as $N \rightarrow \infty$

- Data reweighting
 - Write the change in the posterior expectation as **linear component** + **error**
 - The **linear component** can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
 - As $N \rightarrow \infty$, the linear component provides an arbitrarily good approximation
- High-dimensional problems
 - The linear component is the same order as the error
 - Even for parameters which concentrate, even as $N \rightarrow \infty$
- What should the exchangeable unit be?

Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

Original weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

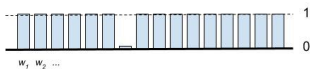
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

Original weights:



Leave-one-out weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

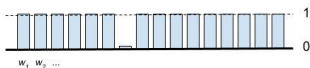
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

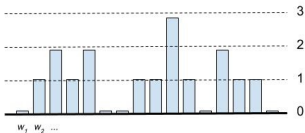
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

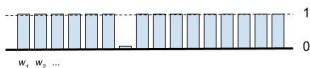
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

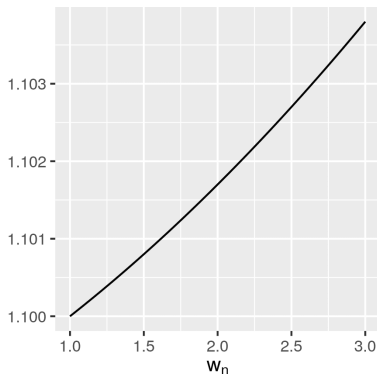
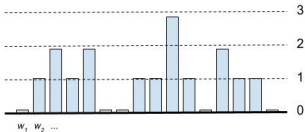
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

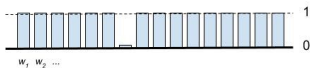
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

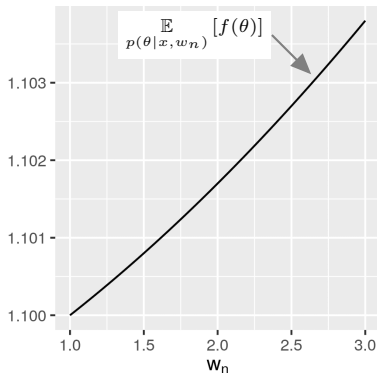
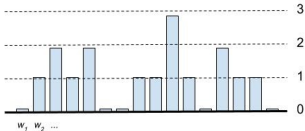
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

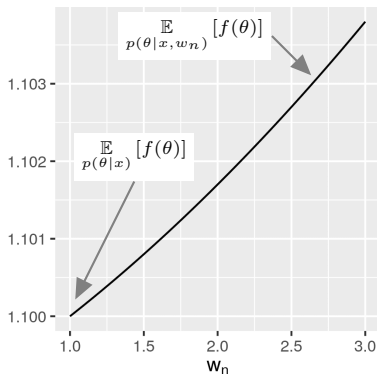
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

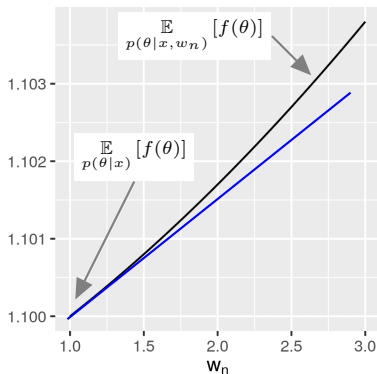
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

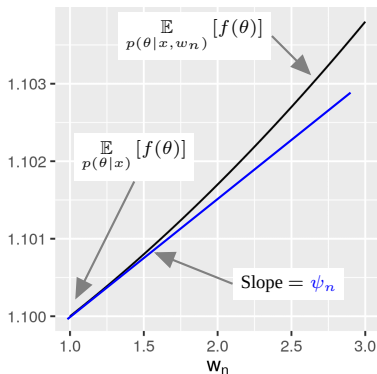
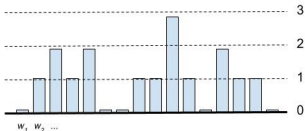
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

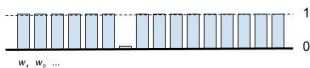
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

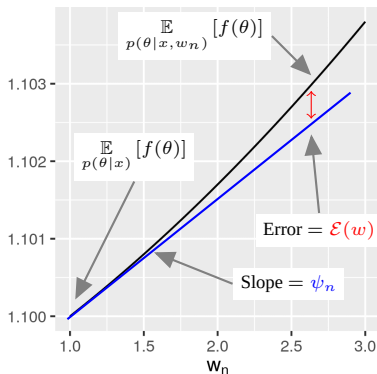
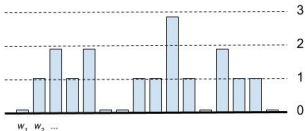
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

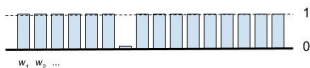
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

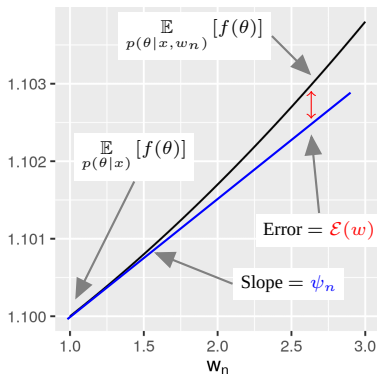
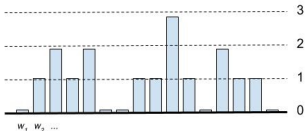
Original weights:



Leave-one-out weights:



Bootstrap weights:



The re-scaled slope $N\psi_n$ is known as the “influence function” at data point x_n .

$$\mathbb{E}_{p(\theta|X,w)}[f(\theta)] - \mathbb{E}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^N \psi_n(w_n - 1) + \mathcal{E}(w)$$

Expressions for the slope and error

How to compute the slopes ψ_n ? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Expressions for the slope and error

How to compute the slopes ψ_n ? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote “posterior–mean zero.” For example, $\bar{f}(\theta) := f(\theta) - \mathbb{E}_{p(\theta|X)} [f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\mathbb{E}_{p(\theta|X)} [\bar{f}(\theta) \bar{\ell}_n(\theta)]}_{\text{Estimatable with MCMC!}} \quad \mathcal{E}(w_n) = \frac{1}{2} \underbrace{\mathbb{E}_{p(\theta|X, \tilde{w}_n)} [\bar{f}(\theta) \bar{\ell}_n(\theta) \bar{\ell}_n(\theta)]}_{\text{Cannot compute directly (don't know } \tilde{w})} (w_n - 1)^2$$

Expressions for the slope and error

How to compute the slopes ψ_n ? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote “posterior–mean zero.” For example, $\bar{f}(\theta) := f(\theta) - \mathbb{E}_{p(\theta|X)} [f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\begin{aligned} \psi_n &= \underbrace{\mathbb{E}_{p(\theta|X)} [\bar{f}(\theta) \bar{\ell}_n(\theta)]}_{\text{Estimatable with MCMC!}} & \mathcal{E}(w_n) &= \frac{1}{2} \underbrace{\mathbb{E}_{p(\theta|X, \tilde{w}_n)} [\bar{f}(\theta) \bar{\ell}_n(\theta) \bar{\ell}_n(\theta)]}_{\text{Cannot compute directly (don't know } \tilde{w})} (w_n - 1)^2 \\ &= O_p(N^{-1}) \text{ under posterior concentration} & &= O_p(N^{-2}) \text{ under posterior concentration} \end{aligned}$$

Expressions for the slope and error

How to compute the slopes ψ_n ? How large is the error $\mathcal{E}(w)$?

For simplicity, for the remainder of the presentation, we will consider a single weight.

$$\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \psi_n (w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote “posterior–mean zero.” For example, $\bar{f}(\theta) := f(\theta) - \mathbb{E}_{p(\theta|X)} [f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\begin{aligned} \psi_n &= \underbrace{\mathbb{E}_{p(\theta|X)} [\bar{f}(\theta) \bar{\ell}_n(\theta)]}_{\text{Estimatable with MCMC!}} & \mathcal{E}(w_n) &= \frac{1}{2} \underbrace{\mathbb{E}_{p(\theta|X, \tilde{w}_n)} [\bar{f}(\theta) \bar{\ell}_n(\theta) \bar{\ell}_n(\theta)]}_{\text{Cannot compute directly (don't know } \tilde{w})} (w_n - 1)^2 \\ &= O_p(N^{-1}) \text{ under posterior concentration} & &= O_p(N^{-2}) \text{ under posterior concentration} \end{aligned}$$

Theorem [Giordano and Broderick, 2023] (paraphrase):

If the posterior $p(\theta|X)$ “concentrates” (e.g. as in the Bernstein–von Mises theorem),^a then

$$w_n \mapsto N \left(\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] \right)$$

becomes linear as $N \rightarrow \infty$, with slope $\lim_{N \rightarrow \infty} \psi_n$.

^aExisting results are sufficient for a *particular weight* [Kass et al., 1990]. Giordano and Broderick [2023] proves that the result holds when averaged over all weights, as needed for variance estimation.

What about when parts of the posterior don't concentrate?

Example: **Generalized linear model with random effects (REs) λ and fixed effect γ .**

Marginally, $p(\lambda|X)$ does not concentrate. Marginally, $p(\gamma|X)$ concentrates.

High dimensional problems

What about when parts of the posterior don't concentrate?

Example: **Generalized linear model with random effects (REs) λ and fixed effect γ .**

Marginally, $p(\lambda|X)$ does not concentrate. Marginally, $p(\gamma|X)$ concentrates.

Does $w_n \mapsto \mathbb{E}_{p(\gamma|X, w_n)} [f(\gamma)] - \mathbb{E}_{p(\gamma|X)} [f(\gamma)]$ become linear as N grows?
(Note $p(\gamma|X)$ *does* concentrate.)

What about when parts of the posterior don't concentrate?

Example: **Generalized linear model with random effects (REs) λ and fixed effect γ .**

Marginally, $p(\lambda|X)$ does not concentrate. Marginally, $p(\gamma|X)$ concentrates.

Does $w_n \mapsto \mathbb{E}_{p(\gamma|X, w_n)} [f(\gamma)] - \mathbb{E}_{p(\gamma|X)} [f(\gamma)]$ become linear as N grows?
(Note $p(\gamma|X)$ *does* concentrate.)

Theorem 5 of Giordano and Broderick [2023] (paraphrase): In general, **no!**

Specifically, if $p(\lambda|X, \gamma)$ does not concentrate, then

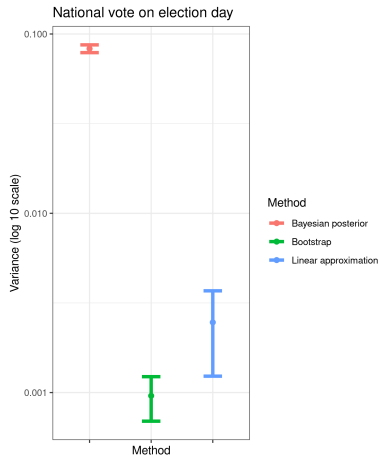
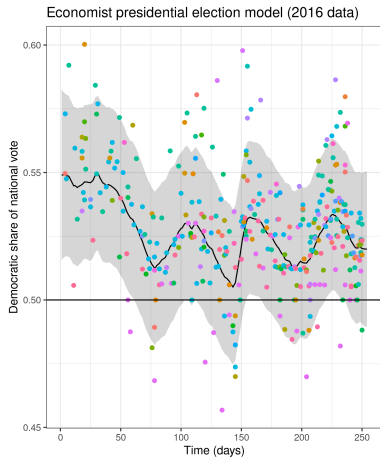
— even if $p(\gamma|X)$ concentrates marginally —

both the slope ψ_n and the error $\mathcal{E}(w_n)$ are $O_p(N^{-1})$, and so

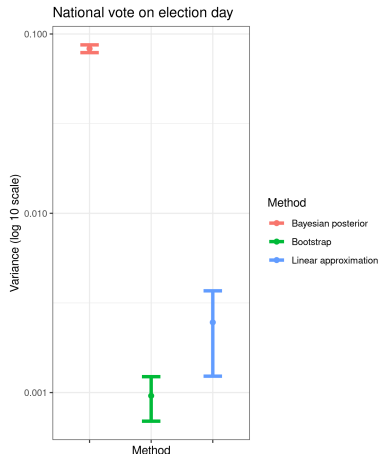
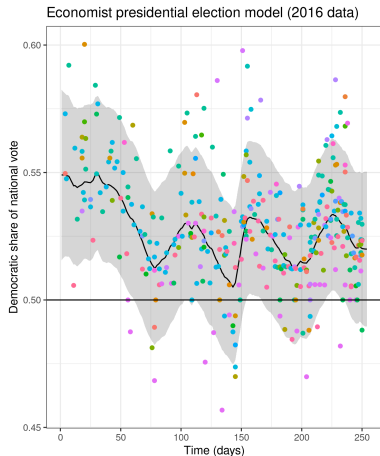
$N \left(\mathbb{E}_{p(\gamma|X, w_n)} [f(\gamma)] - \mathbb{E}_{p(\gamma|X)} [f(\gamma)] \right) = N\psi_n(w_n - 1) + N\mathcal{E}(w_n)$ is nonlinear.

However, $\mathcal{E}(w_n) \rightarrow 0$ as $\text{Cov}_{p(\lambda|X, \gamma)}(\lambda) \rightarrow 0$.

Observations and consequences



Observations and consequences



- We often use models of the form $p(\gamma, \lambda|X)$.
- Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.
... Especially given the linear approximation's huge computational advantage.

Preprint: Giordano and Broderick [2023] (arXiv:2305.06466)

(The preprint focuses on variance estimation, the present results are found in the proofs.)

- A. Gelman and M. Heidemanns. The Economist: Forecasting the US elections., 2020. URL <https://projects.economist.com/us-2020-forecast/president>. Data and model accessed Oct., 2020.
- R. Giordano and T. Broderick. The Bayesian infinitesimal jackknife for variance. *arXiv preprint arXiv:2305.06466*, 2023.
- J. Huggins and J. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.
- R. Kass, L. Tierney, and J. Kadane. The validity of posterior expansions based on Laplace’s method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 1990.

How can we use the approximation?

How can we use the approximation?

Cross validation. Let $w_{(-n)}$ leave out point n , and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathbb{E}_{p(\theta|x, w_{(-n)})} [f(\theta)] \approx \mathbb{E}_{p(\theta|x)} [f(\theta)] - \psi_n$$

How can we use the approximation?

Cross validation. Let $w_{(-n)}$ leave out point n , and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathbb{E}_{p(\theta|x, w_{(-n)})} [f(\theta)] \approx \mathbb{E}_{p(\theta|x)} [f(\theta)] - \psi_n$$

Example: Approximate bootstrap.

Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\begin{aligned} \text{Bootstrap variance} &= \text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|x, w)} [f(\theta)] \right) \\ &\approx \text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|x)} [f(\theta)] + \psi_n(w_n - 1) \right) \\ &= \sum_{n=1}^N \left(\psi_n - \bar{\psi} \right)^2. \end{aligned}$$

How can we use the approximation?

Cross validation. Let $w_{(-n)}$ leave out point n , and loss $f(\theta) = -\ell(x_n|\theta)$.

$$\text{LOO CV loss at point } n = \mathbb{E}_{p(\theta|x, w_{(-n)})} [f(\theta)] \approx \mathbb{E}_{p(\theta|x)} [f(\theta)] - \psi_n$$

Example: Approximate bootstrap.

Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\begin{aligned} \text{Bootstrap variance} &= \text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|x, w)} [f(\theta)] \right) \\ &\approx \text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|x)} [f(\theta)] + \psi_n(w_n - 1) \right) \\ &= \sum_{n=1}^N \left(\psi_n - \bar{\psi} \right)^2. \end{aligned}$$

Influential subsets: Approximate maximum influence perturbation (AMIP).

Let $W_{(-K)}$ denote weights leaving out K points.

$$\max_{w \in W_{(-K)}} \left(\mathbb{E}_{p(\theta|x, w)} [f(\theta)] - \mathbb{E}_{p(\theta|x)} [f(\theta)] \right) \approx - \sum_{n=1}^K \psi_{(n)}.$$

Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^N \text{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^N \text{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

As $N \rightarrow \infty$, $p(\gamma|X)$ concentrates at rate $1/\sqrt{N}$ (Bernstein–von Mises).

$$\Rightarrow N \left(\mathbb{E}_{p(\gamma|X, w_n)} [\gamma] - \mathbb{E}_{p(\gamma|X)} [\gamma] \right) = \psi_n(w_n - 1) + O_p(N^{-1}).$$

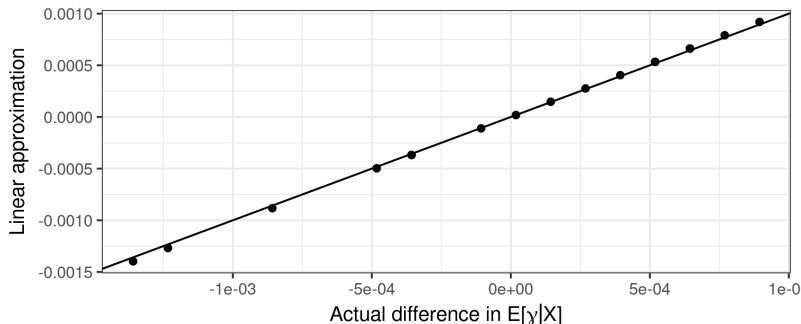
Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^N \text{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

As $N \rightarrow \infty$, $p(\gamma|X)$ concentrates at rate $1/\sqrt{N}$ (Bernstein–von Mises).

$$\Rightarrow N \left(\mathbb{E}_{p(\gamma|X, w_n)}[\gamma] - \mathbb{E}_{p(\gamma|X)}[\gamma] \right) = \psi_n(w_n - 1) + O_p(N^{-1}).$$

Negative Binomial model
leaving out datapoints with $N = 800$



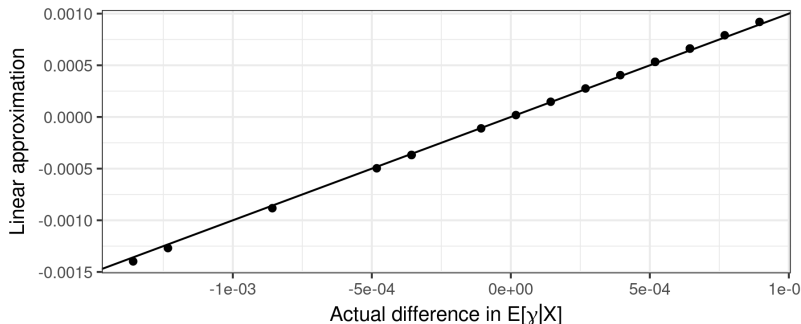
Example: A negative binomial model

Consider $p(X|\gamma) = \prod_{n=1}^N \text{NegativeBinomial}(x_n|\gamma)$. Here, $\theta = \gamma$ is a scalar.

As $N \rightarrow \infty$, $p(\gamma|X)$ concentrates at rate $1/\sqrt{N}$ (Bernstein–von Mises).

$$\Rightarrow N \left(\mathbb{E}_{p(\gamma|X, w_n)}[\gamma] - \mathbb{E}_{p(\gamma|X)}[\gamma] \right) = \psi_n(w_n - 1) + O_p(N^{-1}).$$

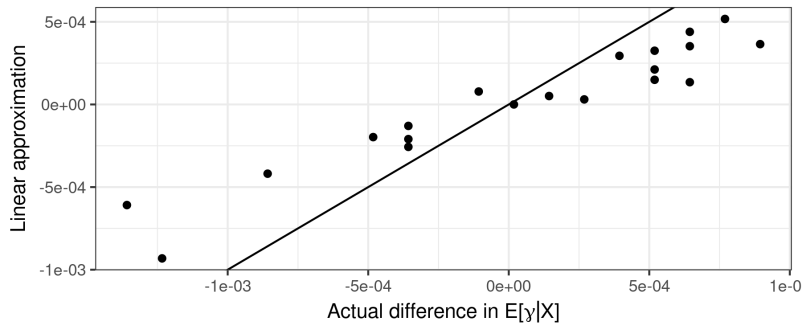
Negative Binomial model
leaving out single datapoints with $N = 800$



Problem: Most computationally hard Bayesian problems don't concentrate.

Example: **Poisson model with random effects (REs) λ and fixed effect γ .**

Poisson random effect model
leaving out single datapoints with $N = 800$



A contradiction?

Negative binomial observations.

Asymptotically linear in w .

Poisson observations with random effects.

Asymptotically non-linear in w .

A contradiction?

Negative binomial observations.

Asymptotically linear in w .

Poisson observations with random effects.

Asymptotically non-linear in w .

With a constant regressor, Gamma REs, and one RE per observation,
these are the same model, with the same $p(\gamma|X)$.

Is $\mathbb{E}_{p(\gamma|X,w)} [\gamma]$ linear in the data weights or not?

A contradiction?

Negative binomial observations.

Asymptotically linear in w .

$$\log p(X|\gamma, w^m) = \sum_{n=1}^N w_n^m \log p(x_n|\gamma)$$

Poisson observations with random effects.

Asymptotically non-linear in w .

$$\log p(X|\gamma, \lambda, w^c) = \sum_{n=1}^N w_n^c \log p(x_n|\lambda, \gamma)$$

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\mathbb{E}_{p(\gamma|X, w)} [\gamma]$ **linear in the data weights** or not?

Trick question! We weight a log likelihood contribution, not a datapoint.

The two weightings are not equivalent in general.

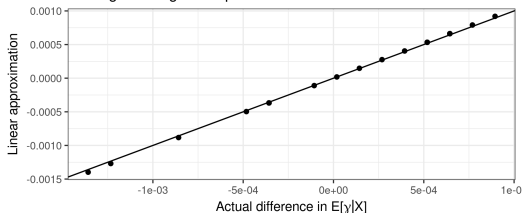
Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n | \gamma)$:

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

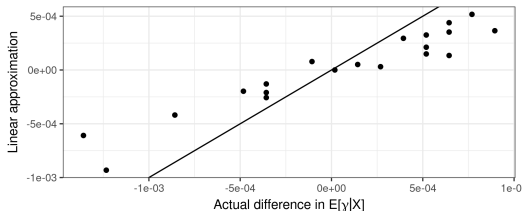
Negative Binomial model
leaving out single datapoints with $N = 800$



Uses $\log p(x_n | \gamma, \lambda)$:

$$\psi_n = \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

Poisson random effect model
leaving out single datapoints with $N = 800$



Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n | \gamma)$:

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

Not computable from

$$\gamma, \lambda \sim p(\gamma, \lambda | X)$$

in general.

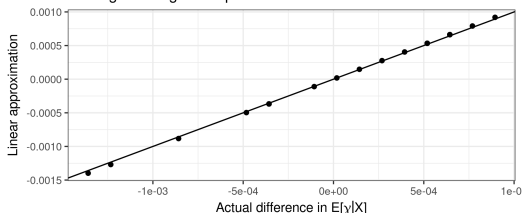
Uses $\log p(x_n | \gamma, \lambda)$:

$$\psi_n = \mathbb{E}_{p(\gamma, \lambda | X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

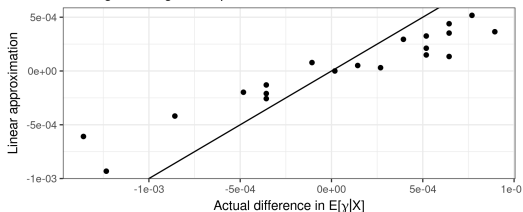
Computable from

$$\gamma, \lambda \sim p(\gamma, \lambda | X).$$

Negative Binomial model
leaving out single datapoints with $N = 800$



Poisson random effect model
leaving out single datapoints with $N = 800$



Experimental results

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n | \gamma)$:

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

Not computable from

$$\gamma, \lambda \sim p(\gamma, \lambda | X)$$

in general.

Uses $\log p(x_n | \gamma, \lambda)$:

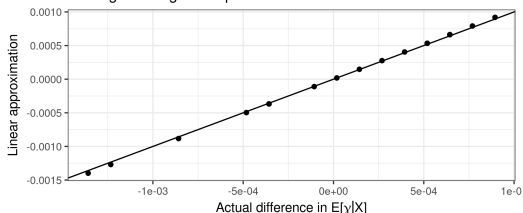
$$\psi_n = \mathbb{E}_{p(\gamma, \lambda | X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

Computable from

$$\gamma, \lambda \sim p(\gamma, \lambda | X).$$

May still be useful when $p(\lambda | X)$ is *somewhat* concentrated.

Negative Binomial model
leaving out single datapoints with $N = 800$



Poisson random effect model
leaving out single datapoints with $N = 800$

